

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re U.S. Patent Application of)
)
KOIKE et al.)
)
Application Number: To be Assigned)
)
Filed: Concurrently Herewith)
)
For: NETWORK DRAWING SYSTEM AND)
NETWORK DRAWING METHOD)
)
)
ATTORNEY DOCKET NO. ASAM.0101)

Honorable Assistant Commissioner
for Patents
Washington, D.C. 20231

**REQUEST FOR PRIORITY
UNDER 35 U.S.C. § 119
AND THE INTERNATIONAL CONVENTION**

Sir:

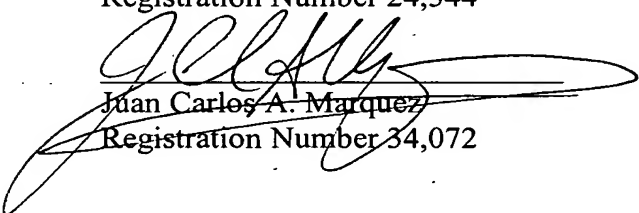
In the matter of the above-captioned application for a United States patent, notice is hereby given that the Applicant claims the priority date of October 14, 2003, the filing date of the corresponding Japanese patent application 2003-353097.

A certified copy of Japanese patent application 2003-353097 is being submitted herewith.

Acknowledgment of receipt of the certified copy is respectfully requested in due course.

Respectfully submitted,

Stanley P. Fisher
Registration Number 24,344


Juan Carlos A. Marquez
Registration Number 34,072

REED SMITH LLP
3110 Fairview Park Drive
Suite 1400
Falls Church, Virginia 22042
(703) 641-4200
January 29, 2004

日 本 国 特 許 庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日 2 0 0 3 年 1 0 月 1 4 日
Date of Application:

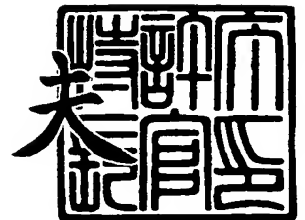
出 願 番 号 特 願 2 0 0 3 - 3 5 3 0 9 7
Application Number:
[ST. 10/C] : [J P 2 0 0 3 - 3 5 3 0 9 7]

出 願 人 株 式 会 社 日 立 製 作 所
Applicant(s):

2 0 0 3 年 1 2 月 2 5 日

特許庁長官
Commissioner,
Japan Patent Office

今 井 康 夫



出証番号 出証特 2 0 0 3 - 3 1 0 7 4 2 6

【書類名】 特許願
【整理番号】 H03009871A
【あて先】 特許庁長官 殿
【国際特許分類】 G06F 17/30
【発明者】
 【住所又は居所】 東京都国分寺市東恋ヶ窪一丁目 2 8 0 番地 株式会社日立製作所
 中央研究所内
 【氏名】 小池 麻子
【発明者】
 【住所又は居所】 東京都国分寺市東恋ヶ窪一丁目 2 8 0 番地 株式会社日立製作所
 中央研究所内
 【氏名】 丹羽 芳樹
【発明者】
 【住所又は居所】 東京都江東区中島一丁目 3 - 1 6 - 6 0 9
 【氏名】 高木 利久
【特許出願人】
 【識別番号】 000005108
 【氏名又は名称】 株式会社日立製作所
【代理人】
 【識別番号】 100075096
 【弁理士】
 【氏名又は名称】 作田 康夫
 【電話番号】 03-3212-1111
【手数料の表示】
 【予納台帳番号】 013088
 【納付金額】 21,000円
【提出物件の目録】
 【物件名】 特許請求の範囲 1
 【物件名】 明細書 1
 【物件名】 図面 1
 【物件名】 要約書 1

【書類名】 特許請求の範囲**【請求項 1】**

第 1 のカテゴリーに属する第 1 のクエリーを指定する第 1 の入力部と、
第 2 のカテゴリーに属する第 2 のクエリーを指定する第 2 の入力部と、
前記第 1 のカテゴリーと第 2 のカテゴリーとを含む第 3 のカテゴリーに属する用語間の
関連度を、複数組テーブル状に記憶したデータ格納手段と、
前記データ格納手段に記憶されたテーブルを用い、入力された前記第 1 のクエリーと前
記第 2 のクエリーとを、複数の用語を介して関連づける計算手段と、
前記計算手段の計算結果に基づいて、前記第 1 のクエリーと前記第 2 のクエリーとを、
複数の用語を介して連結した用語のネットワークを、画面表示する表示手段とを有するこ
とを特徴とする画面表示システム。

【請求項 2】

前記画面表示システムは、更に描画条件を指定する第 3 の入力部を有し、
前記ネットワークは、前記第 3 の条件に基づき、表示されることを特徴とする請求項 1
記載の画面表示システム。

【請求項 3】

前記データ格納手段は、更に前記用語の付随情報を記録していることを特徴とする請求
項 1 記載の画面表示システム。

【請求項 4】

前記第 1 のクエリーと前記第 2 のクエリーの少なくとも何れかは、複数であることを特
徴とする請求項 1 記載の画面表示システム。

【請求項 5】

前記第 1 のクエリーと前記第 2 のクエリーとを繋ぐルートの中で、前記用語間の関連度
が最も高いものを強調線で繋ぎ、表示することを特徴とする請求項 1 記載の画面表示シ
ステム。

【請求項 6】

前記第 1 のカテゴリーは、疾患名、症状、蛋白質名、遺伝子名、化合物名、遺伝子の機
能、蛋白質の機能の少なくとも何れか、

前記第 2 のカテゴリーは、化合物名、蛋白質名、遺伝子名の少なくとも何れかであるこ
とを特徴とする請求項 1 記載の画面表示システム。

【請求項 7】

前記用語間の関連性は、用語間の共起、もしくはフレーズパターンにより抽出されるこ
とを特徴とする請求項 1 記載の画面表示システム。

【請求項 8】

前記第 3 の入力部の設定を変更することにより、インタラクティブに用語間のネットワ
ークを再表示することを特徴とする請求項 1 記載の画面表示システム。

【請求項 9】

前記第 3 の入力部の設定を変更することにより、インタラクティブに、用語間の連結も
しくは用語自身の追加・削除の編集を行えることを特徴とする請求項 1 記載の画面表示シ
ステム。

【請求項 10】

前記描画表示システムは、更に、前記第 1、前記第 2 の入力部から入力された少なくと
も一方のクエリーを、代表的な用語に変換する同義語辞書を有することを特徴とするこ
とを特徴とする請求項 1 記載の画面表示システム。

【請求項 11】

前記用語間の関連性が、前記画面上に合わせて表示されることを特徴とする請求項 1 記
載の画面表示システム。

【請求項 12】

前記用語が階層性を有するときに、上位の階層の用語に上げて表示することを特徴とす
る請求項 1 記載の描画表示システム。

【請求項 13】

前記第2のカテゴリーは遺伝子名であり、前記遺伝子名を前記画面上の横軸に表示し、更に、ロッドスコアを、前記横軸の遺伝子ごと、もしくは染色体位置情報と共に合わせて表示されることを特徴とする請求項1記載の画面表示システム。

【請求項 14】

前記用語間の関連性を、遺伝子クラスタリング結果と合わせて表示することを特徴とする請求項1記載の画面表示システム。

【請求項 15】

前記ネットワークの表示結果と前記遺伝子クラスタリングの結果が対応しないときに、前記対応しない第1のクエリーと前記第2のクエリーを結ぶルートを強調線で繋ぎ、表示することを特徴とする請求項1記載の画面表示システム。

【請求項 16】

第1の入力部に、第1のカテゴリーに属する第1のクエリーが入力されるステップと、第2の入力部に、第2のカテゴリーに属する第2のクエリーが入力されるステップと、前記第1のカテゴリーと第2のカテゴリーとを含む第3のカテゴリーに属する用語間の関連度を、複数組テーブル状に記憶したデータ格納手段を用い、入力された前記第1のクエリーと前記第2のクエリーとを、複数の用語を介して関連づけるステップと、

前記関連づけた結果に基づいて、表示手段に、前記第1のクエリーと前記第2のクエリーとを、複数の用語を介して連結した用語のネットワークを、画面表示するステップとを有することを特徴とする画面表示方法。

【請求項 17】

前記データ格納手段に、インターネットを介して接続することを特徴とする請求項16記載の画面表示方法。

【書類名】明細書**【発明の名称】画面表示システム及び画面表示方法****【技術分野】****【0001】**

本発明は、データベースに蓄積されたキーワード、データ等の関連性情報から、用語間のネットワークの構築を支援する画面表示システム及びその方法に関する。

【背景技術】**【0002】**

一般に、情報検索の分野では、キーワード等の検索キーを元に、そのキーワードと関連性の高い検索結果を抽出し、その結果を画面に表示する。例えば、WO01/020535号には、複数のデータベースを用いて、様々な検索の仕方により、生物学的データを検索することが記載されている。

【0003】

一方、近年の分子生物学の急激な進歩とゲノムの完全解読によって、遺伝子や疾患に関する情報が増大している。しかしながら、文献に蓄えられている知識と新たに得られた実験結果とは独立に扱われており、自動的に両者を統合する手段が殆どなかった。特に、ゲノムの完全解読によって進展が望まれている連鎖解析、関連解析の分野では皆無であった。従って、連鎖解析、関連解析の分野で疾患遺伝子の候補となる染色体の部位が限定されたとしても、その候補範囲に存在する遺伝子数が100を下らないことも多く、通常研究者はその候補となる遺伝子がどのような機能に関するものか一つ一つ文献を読んで疾患遺伝子を検討・推定し、次の実験ステップを選定していた。また、DNA-アレイ、Protein-アレイの発現情報のクラスタリングにおいても、クラスタリングの妥当性は、クラスタリング内の遺伝子がかつて関連性を指摘された遺伝子か否か、研究者が文献を読んで判定していた。

【0004】

【特許文献1】WO01/020535号

【発明の開示】**【発明が解決しようとする課題】****【0005】**

今後も研究の進展に伴い、特にバイオの分野では蛋白質相互作用情報、発現情報、転写因子情報等に関する各種実験が多数行われ、その膨大な結果が蓄積されていくと考えられる。従って、研究者は新たな実験によって得られたデータと既に得られている情報との関係を鑑みて生物学的な知見を得るためには、文献の検索等に膨大なエネルギーを費やす必要がある。

【0006】

本発明の目的は、文献の検索効率を高め、検索者が容易に、用語間の関係情報を得られるようにすることにある。

【課題を解決するための手段】**【0007】**

上記課題を解決するため、本発明は、ユーザが関係を知りたい用語群1と用語群2を指定することにより、予め蓄積した用語間の関係、もしくはインターネットを介して動的にアクセスして得られる用語間の関係を用いて、いかにして用語群1と2が関連付けられるか表示することにより、研究者が文献を逐一読むことなく、実験的に得られた情報とこれらの情報を重ね合わせ、新たな生物学知見を得ることができる。

【0008】

具体的には、以下の構成とする。第1のカテゴリーに属する第1のクエリーを指定する第1の入力部と、第2のカテゴリーに属する第2のクエリーを指定する第2の入力部と、前記第2のカテゴリーと第1のカテゴリーを含む第3のカテゴリーに属する用語間の関連度とその付随情報を複数組テーブル状に記憶したデータ格納手段と、前記データ格納手段に記憶されたテーブルを用い、入力された前記第1のクエリーと前記第2のクエリーとを

、複数の用語を介して関連づける計算手段と、前記計算手段の計算結果に基づいて、前記第1のクエリーと前記第2のクエリーとを、複数の用語を介して、連結するように、ネットワーク上に画面表示する表示手段とを有することを特徴とする画面表示するシステム、である。更に、検索条件を指定する第3の入力部を設けても良い。

【0009】

例えば、第一のカテゴリには、化合物、疾患名、疾患症状、蛋白質／遺伝子名などが考えられ、第2のカテゴリには化合物、蛋白質／遺伝子名などが考えられるが、ユーザが興味を持っている2つの用語群の関係であればこれに限定されない。生物学カテゴリ以外では、例えば、第1のカテゴリは機器の故障症状、第2のカテゴリに機器の機種を入れ、第1と第2のカテゴリを故障原因となる名詞句で結ぶことにより、機種ごとの故障症状と故障原因の関係をざっくり見ることができる。また第1のカテゴリに政治家の名前、第2のカテゴリに省庁名を入れてどのような関連性があるか見ることも可能だし（この場合はネットワークを結ぶ用語は全ての名詞句が対照となる）、第1のカテゴリに外国の都市名、第2のカテゴリに日本の都市名を入れ、都市同志の共通点で結ぶことも可能である。また、第1のカテゴリは、特許文献におけるキーワード、第2のカテゴリは、論文におけるキーワードなどとしても良い。

【0010】

ここでの用語間の関係は、Web上に公開されているデータや文献を解析して得られるものの全てを含む。文献からのデータの抽出は、人が読んだ上で抽出するもの、自然言語処理等の機械処理で自動的に抽出するものを含む。自然言語処理による用語間の関係は、主に、共起、フレーズパターンなどによって抽出される。

【0011】

用語間のネットワークは、上述の用語間の情報の重み（用語間の関連性）を考慮して描画することになる。2点間の用語の最短距離はダイクストラ法や日程計画法によって記述される。ここでの距離とは、用語間の関連度が高い最短距離が高いほど用語間の距離が短くなる関数で定義されるものである。必ずしも最も重要な用語のパスになるとも限らないので、高得点側からの候補をいくつか表示することが望ましい。

【0012】

また、用語間の関連度が尤も高いパスを強調線で繋ぎ表示することもできる。

【0013】

更に、最短距離の計算はダイクストラ法等では、すべての点の距離を計算すると計算時間がかかることから、ユーザの指定により、閾値よりも遠い用語間の距離を計算しないように適宜、枝刈りすることが望ましい。この閾値の指定は、第3の入力部により行うことができる。また、対象とする用語数が多い場合は、予め第1のクエリーと第2のクエリーを結ぶ最長ステップ数（間に入る用語の数）を第3の入力部により限定しておく、計算時間が短くて済む。

【0014】

発明によれば、例えば、連鎖解析によってロッドスコアを得、疾患遺伝子の候補となる遺伝子の領域が定められている場合、この中から既知の知識を総括して疾患遺伝子として最も尤もらしい遺伝子もしくは遺伝子群を提供することができる。

【0015】

また、発明によれば、DNA アレイやProtein アレイの遺伝子／蛋白質クラスタリング結果と同時に用語のネットワークを表示することにより、実験によるノイズと思われるクラスタを構成している遺伝子／蛋白質を提示することができる。

【0016】

また、このシステムにおいては、用語をつなぐエッジをクリックすることにより、用語間の関連性を示すデータの出展元となる雑誌名、情報を抽出したセンテンス、アブストラクト、データベース名を提示することができる。また、ノードをクリックすることにより、それぞれの用語の付随情報、例えば、用語が蛋白質の場合は細胞内局在性や、発現情報が閲覧可能である。

【0017】

発明によれば、用語として用いる用語群がgene ontology(<http://www.geneontology.org/>)やfamily名のように階層性を有するものである場合、上位の階層でネットワークを描画することにより、ネットワークを簡潔に、もしくは、出現頻度が低く統計的に不確かな用語間の関係を考慮した表示に、もしくは、ノード（用語）を連結する条件を緩和したネットワークの表示を実行することができる。

【0018】

一方、このシステムにおいては、着目する生物種に関する遺伝子／蛋白質の関連性を示すデータが少ない場合、配列解析によって、他の生物種と着目する生物種間のオルソログもしくは類似遺伝子／蛋白質を結びつけることにより、他の生物種の情報を使って、ネットワークを構築することもできる。具体的には、他の生物種の情報と配列相同性やドメイン構成情報などを用いる。

【0019】

また、発明によれば、画面表示システムにおける編集機能の追加により、不適当な用語間の連結（エッジ）、もしくは用語自身を除去し、また、不足と思われる用語間の連結もしくは用語自身を追加し、インタラクティブにネットワークを再構築することができる。

【発明の効果】**【0020】**

上記で説明したように、本発明においては、用語間の情報を2項関係とその付随情報として蓄積した情報を用い、クエリー1とクエリー2の間をつなぐ用語のネットワークを表示することにより、用語間の関係を知ることができる。これにより、関連性がないとされていた概念（用語）に関しても、関連性を見出すことが容易になり、検索の利便性が高まる。

【発明を実施するための最良の形態】**【0021】**

以下、図面を参照して本発明の実施形態の一例を詳細に説明する。本発明は、その本旨を越えない限り、以下の実施例に限定されるものではない。

【実施例1】**【0022】**

図1は本発明による用語ネットワークシステムの実施形態の一例を示すシステム構成図である。用語ネットワークシステムは入出力装置／クエリー入力部1、中央計算処理装置2、入出力装置／表示部3、データ格納装置4から構成される。必要に応じて、実験データ入力装置5を追加する。また、使用する用語の同義語の問題を解消するために、クエリーを代表的な用語に変換する辞書6を備えているとより望ましい。

【0023】

このシステムの利用方法を図2に示す。文献や各種データベースから人手または自動的に2項／多項関係を抽出し、その関連度の情報と共に蓄積したテーブルand/or 共起を計算するための用語とその文献情報を蓄積したテーブルと、その用語もしくは用語間の付随情報を図1-データ格納装置4に蓄積しておき、図1-入力部1で指定する用語群クエリー1(図2-ステップ1)とクエリー2(図2-ステップ3)を検索条件(図1-入力部13、図2-ステップ4)で結合スコアが高くなる用語のネットワークを図1-中央計算装置2で構築し、図1-表示部3で表示する(図2-ステップ5)。ユーザの目的に応じクエリー1とクエリー2を結ぶネットワークのうち、関連度のスコアが最も高くなるネットワークを強調表示する(図2-ステップ6)。連鎖解析、関連解析の実験結果がある場合は、図1-入力部1の入力部3における検索条件で、実験結果から候補となる遺伝子領域を指定することにより(図2-ステップ2)、クエリー2はその領域の遺伝子名が自動的に指定される(図2-ステップ3)。この場合、クエリー1に疾患名もしくは疾患名と症状などが入るが、これに限定されない。更に、クエリー1とクエリー2を結ぶ用語ネットワークのスコアを計算し、過去の情報における各遺伝子と疾患遺伝子の関連度と実験結果のロッドスコアと並べて表示することにより、候補遺伝子領域を提示する(図2-ステップ7)。また、疾

患遺伝子と疾患名を結ぶスコアの高いネットワークを構成する用語を合わせて表示することにより、具体的な関連性を明確にする(図2-ステップ8)。アレイデータを使う場合は、ステップ2で対象とする遺伝子群(クエリー2となる)と発現情報からクラスタリングした結果を入力する。上述と同様の手順を踏み、ステップ9において、用語のネットワークを用いクラスタリングした結果の中からノイズの候補となる遺伝子を提示する。

【0024】

図1におけるデータベース格納装置4には、予め文献から人手または文型を使って自動抽出した相互作用や遺伝子・蛋白質などの機能情報等を2項関係とし(図1-42、43)、また、その他のデータベースから取り出した相互作用情報や構造機能情報(44)を蓄積してある。文献中の用語の共起情報を使う場合は、用語と用語が出現した文献及び文献中の用語位置をテーブルとして蓄積しておく(41)。対象となる用語が少ない場合は、予めすべての用語ペアの共起を計算して2項関係の重みとしたテーブルにしても構わない。これらの情報を自動抽出する場合の流れは図3に示す。対象データとしては、各種科学専門雑誌やNCBI (<http://www.nlm.ncbi.nih.gov>) PUBMED-abstract, PUBMED-centralなどに登録された雑誌となる。対象論文は、PUBMEDを使用する場合は予め、mesh termを使い対照とする生物種のアブストラクト/論文のみに絞っておいたほうが、他の生物種の情報が混在せず、望ましい。なお、データ格納装置には、インターネットを介して接続するようにしても良い。

【0025】

続いて、用語間の関係を抽出する方法について説明する。ネットワークを構成する用語群としては、遺伝子/蛋白質名、化合物名、gene ontology, UMLS (Unified Medical Language System), SNOMED (International: The Systematized Nomenclature of Medicine), Mesh (Medical Subject Headings) など、人手でコントロールしてある用語集/辞書、もしくはその組み合わせが望ましいが、文中に現れるすべての名詞句等を用語として使用しても構わない。また、文中に現れる全ての名詞句のうち、新聞等の対照とする他のコーパス中に現れる名詞句の使用頻度よりも高い名詞句のみを使用する用語の対照にしてもよい。或いは、近隣の語基との相互情報量の利用(例えば、Shimohata et al ACL 1997)、C-value method (Maynard and Ananiadou, TKE 1999) 等によって自動的に用語セットを対象となる文献から抽出してもよい。また、ターゲットとなる用語の部分セットとそれらが出現しやすい局所文脈とを用いて残りの用語(及び局所文脈)を対象となる文献から自動抽出するBooststrap法を用いて用語セットを作成してもよい。(例えば、Agichtein et al. 2001 2001 ACM SIGMOD International Conference on Management of Data)

これらの用語はできる限り、辞書などを使って、もしくは必要に応じて辞書を構築して同義語、同音異義語等の解決をしていることが望ましい。

【0026】

フレーズパターン(文型)による抽出においては構文解析や係り受け解析を行なうことにより noun phrase bracketing を行った後、挿入句、等位接続等の文の構造解析を行ない、名詞句 activate 名詞句、名詞句 interact with 名詞句、名詞句 inhibit 名詞句、interaction between 名詞句 and 名詞句などにより、名詞句に目的とする用語が入っているかどうかチェックすることにより用語間の関係を抽出する。例えば A domain of protein-1 is activated by B domain of protein-2 という文によって protein-2 が protein-1 を活性化するという情報を自動的に抽出できる。2つの用語間の関係の強さは、その関係の記述頻度だけでなく、関係抽出時の確からしさを用語間の単語の距離や、文法的な複雑さ等、(例えば、抽出した蛋白質名が名詞句の中で前置詞もしくは特定の用語の後ろに位置しているか否かなど)を使用して表すことができる。構文解析や係り受け解析の前に、その解析精度を上げるために必要に応じて対象とする用語をIDに変換したり、複数の単語からなる専門用語を noun bracketing する前処理を行ってもよい。

【0027】

用語間の関係抽出は、様々な方法があり、これに限定されない。フレーズパターンによって情報抽出をした例を図4に示す。この図においては、2項間の関連性を抽出時の信頼

度として記している。ここで、信頼度は関係を抽出したフレーズパターンの種類と遺伝子が含まれる名詞句の前に特定の前置詞がくるか否かによって、求められる。2項関係自身に付随情報として情報の重みや実験方法、情報の出展元の文献情報、等が付けられている。

【0028】

用語間の共起関係を求める場合は、用語間の相互情報量などを使うことができるが、様々な方法がありこれに限定されない。用語間の相互情報量は F_{ab} = 用語Aと用語Bが共出現する単位テキストの数, F_a = 用語Aが出現する単位テキストの数, F_b = 用語Bが出現する単位テキストの数, N = 単位テキストの総数, としたとき, $\log(F_{ab} * N / F_a / F_b)$ で求められる。また、この値と F_{ab} との積を掛けた $F_{ab} \log(F_{ab} * N / F_a / F_b)$ (エントロピーゲイン) も有効である。また、更に、PHGS (N 、 n 、 K 、 k) を N 個の玉、内 K 個が赤球である袋から無作為に n 個を取り出したときに赤球が k 個以上含まれる確率値とすると、 $-\log(\text{PHGS}(N, F_a, F_b, F_{ab}))$ という値やこれを対称化した $-\log(\text{PHGS}(N, F_a, F_b, F_{ab})) - \log(\text{PHGS}(N, F_b, F_a, F_{ab}))$ など有効な共起度の尺度である。単位テキストとしては、文書全体、章、節、パラグラフ、文、一つの用言が支配する範囲内(単文)などの構造上の単位や構造とは無関係に一定の単語数の範囲内というような設置をすることができる。2つの用語間の関係の強さは、これらの式によって一意に決めることができる。

【0029】

用語間の共起関係は予め計算してテーブルにしておいてもよいが、図1における中央計算処理装置によって、動的に計算してもよい。用語とその文献情報の例を図5に示す。

【0030】

関係を構成している用語にも付随情報として遺伝子の発現情報や細胞内局在情報が付けられる。付随情報として細胞内局在性や配列相同性を示すE-value(同じ類似度の配列がそのデータベース内で偶然現れる何本現れるかを示す期待値)などの例を図6に、文献情報を図7に示す。用語それ自身が階層構造となる概念である場合は、その情報もデータ格納装置4に付加される。図8に例を示す。発現情報や局在情報などの付随情報は関係を抽出した文献だけからでなく他の実験結果の情報を蓄積してもよい。ここでの情報の重みとは、共起を使ったときは出現頻度、フレーズを使ったときには上述の抽出時の信頼度と出現回数、人手で抽出した場合は全てが一律に1とするなどであり、用語間の関連度を示す。複数の異なる抽出データを使う場合は、データ間で整合性がとれるように必要に応じて正規化する。以下、この重みをスコアと呼ぶ。また、Web等を介して外部のデータベースから用語間の関係を動的に蓄積しつつ以下の計算を行うことも可能である。

【0031】

クエリーの入力部1には、クエリー1とクエリー2の2つのクエリー群とその他の検索・描画条件設定部門からなる。具体的な入力装置の画面を図14に示す。検索画面には、クエリー1, 2を入力するためのウインドウ81, 82と、検索条件を指定する第3の入力部であるウインドウ83がある。第3の入力部では、用語間の距離を指定、高得点側からの候補数の指定、関連度に応じて強調線を利用することを指定、用語間の関係を記憶したデータセットの種類を指定、データの信頼度に関する情報の指定、クエリー1, 2を結ぶ最長ステップ数の指定等を行うことができる。クエリー1, 2においてスペースのある連語を入力する場合は'や"'で囲む、もしくはアンダーバー()などで結び、複数のクエリーを指定する場合は、スペースで区切りそれらのクエリーを入力することが可能である。入力方式は様々な方法が考えられるので、これに限定されない。

【0032】

クエリー1とクエリー2は主に遺伝子/蛋白質/化合物及びその機能、疾患名、症状等ユーザの要求に応じて指定する。クエリー1, クエリー2とも1つ以上の用語からなる。中央計算処理装置においてクエリー1とクエリー2に属する用語を用語間の関連度を示すスコアを使い、スコア総和/(エッジ数^{1.1})、もしくは、その他のスコアとエッジから

なる関数により、クエリー1とクエリー2を結ぶ高スコアの用語ネットワーク候補をダイクストラ法や日程計画法等によって計算する。必ずしも最も高い得点が最良のネットワークではないことからユーザが入力部3で指定する検索条件における候補数を同時に計算する。ネットワークの計算対象とするデータセットやそのデータセットが階層的な概念からなるときに、ユーザが図1-入力部3で指定する上位階層において用語ネットワークを書くことができる。例を図9, 10に示す。図9においては、複雑だったネットワークが、上位概念(用語)で描画することにより理解しやすい図10になる。上位概念の描画は、描画条件の緩和を意味する場合もある。例えば、図9において、RRASとRAS1、及びRAF1とMAP2K1のみの関連性が指摘されていたときは、RRASからMAP2K1の間の関連性は抽出されない。しかし、上位の概念においては、この情報においてRASとRAFが、RAFとMAP2Kが関連付けられるのでRASとMAP2Kが関連付けられる。

【0033】

更に、ユーザは図1-入力部3によって描画の条件設定をインタラクティブに行える。例えば、用語ネットに使用するデータセットの種類や、データの信頼度によるスクリーニングなどの指定を行うことができる。スクリーニングの例としては、大量実験手法である yeast-two hybridや質量分析の結果は信頼性が低いので除く、もしくは、impact factor が低い学術誌に記述されているものは除くなどがある。ここでの impact factor は Institute for Scientific Information が算出している値だけではなく、他の団体・機関が算出する値も含む。また、文献からの自動抽出エラーを除くために、フレーズパターンによる自動抽出の場合はある一定数以上の文献情報があることなどを条件にしてもよい。更に、2項関係情報のうち蛋白質-蛋白質/DNA/RNA間の相互作用情報を使う場合は、実験エラーを除くために、細胞内局在性が著しく異なる2項間の関係、(例えば片方が核内蛋白質で他方がミトコンドリア内の蛋白質など)は除外してネットワークを構築することも可能である。また、特定の細胞で発現している遺伝子/蛋白質のみを使用してネットワークを構築することも可能である。例えば、H. sapiens の BCL-2 と相互作用している遺伝子は現在10以上あるが特定の組織に限ればその数は減少する。着目する組織を lymphocyte とした場合、BCL-2 相互作用する遺伝子として知られる PSEN1 は少なくとも現在は lymphocyte での発現情報が報告されていないので、BCL-2 と PSEN1 の関連性は用語のネットワーク構築には使用されないことになる。

【0034】

更に、ユーザは、画面の編集機能によって、描画したネットワークのうち、不適と思われるエッジ(用語と用語を結ぶ線)もしくは用語自身を除去したり、その反対にエッジの追加や用語自身の追加を行い、ネットワークを再計算することも可能である。

【0035】

また、着目する生物種に関して、遺伝子・蛋白質の関連情報が少ない場合は、他の生物種の情報と配列類似性を用いて、同様に用語のネットワークを構築することも可能である。例えば、図6の配列相同性に相当する情報を用いて、S. cerevisiae から C. elegans のネットワークを構築することも可能である。また、E-value のトップスコアもしくは閾値などを使って対応する遺伝子を見つけるだけでなく、目的に応じてドメイン情報などの緩い条件を使って生物種間で蛋白質/遺伝子を対応させてもかまわない。例えば、図4と図6においては、S. cerevisiae における STE20 と STE11 の関連性はトップスコアを用いると C. elegans において GCE000836(mig-15)と GCE000678(kin-18)の関連性に射影される。

【0036】

【非特許文献1】Shimohata et al ACL 476-481, 1997

【0037】

【非特許文献2】Maynard and Ananiadou, TKE 212-221, 1999

【非特許文献3】Agichtein et al. 2001 ACM SIGMOD International Conference on Management of Data

【実施例2】

【0038】

連鎖解析の結果への利用に関しては、idiopathic hypogonadotropic hypogonadismの原因遺伝子が染色体19p13.2に存在すると考えられた例を図11、12で説明する。連鎖解析の結果はJS. Acierno et al. The Journal of Clinical Endocrinology and Metabolism 88(6), 2947-2950である。図11において $q=0.05$ の解析結果を実線で示す。(handbook of statistical genetics, edited by DJ. Balding et al. Wiley, England 2001に連鎖解析に q の意味は参照) 点線がクエリー1をidiopathic hypogonadotropic hypogonadismとして、1980-2003のhumanのPUBMEDに登録されているアブストラクトの名詞句の共起関係から計算した用語間のネットワークのスコアを点線で示す。この結果、両者がピークを持つ、0.2-0.4 Mb位置が図2-ステップ7 (図11の矢印に挟まれた部分) で提示されることになる (ここでは、0.2 Mbをmarker rs7815の位置とした)。また、図2-ステップ8における用語のネットワークの例は図12に示す。図12では、第2のカテゴリーが遺伝子名であり、画面の横軸に遺伝子名を表示し、ネットワーク状に、第1、第2のクエリーとを複数の用語を介して画面表示した例である。このネットワーク表示に、更に、ロッドスコアを合わせて表示し、ロッドスコアは、横軸の遺伝子ごと、もしくは染色体位置情報と共に合わせて表示した。ここでは、ステップ7において選択したLod scoreと文献情報から尤もらしいと思われる領域0.2-0.4 Mbを表示した例である。尤もスコアが高くなる用語ネットを太線で表示している。なお、ロッドスコアについては、Onda et al. Stroke, 34(7), 1640-1644, 2003に、詳細が記載されている。

【0039】

そして、この実施例では、用語を繋ぐ線分をクリックすることで、データ格納装置に格納されたデータの情報、例えば、用語間の関連性を示すデータの出典もとなる雑誌名、センテンス、アブストラクト、データベース名等を表示することができる。また、ノードをクリックすることで、用語の付随情報を表示することができる。これは、インターネットにリンクさせ、情報を抽出しても良い。

【0040】

【非特許文献4】JS. Acierno et al. The Journal of Clinical Endocrinology and Metabolism 88(6), 2947-2950, 2003

【非特許文献5】Onda et al. Stroke 34(7), 1640-1644, 2003

【実施例3】

【0041】

DNA-アレイの発現データを使った場合の用語ネットワークの例を図13に示す。図2ではステップ9に相当する。図13の下(72)は、実験結果からクラスタリングした例を示している。クエリー1はgene ontologyのbiological processに分類される用語すべてを入れ、クエリー2に発現データで同じグループにクラスタリングされた遺伝子名を入れた場合の用語のネットワークが記述されている。この例では、有意なネットワークを示さないクエリー1の用語は表示されない(71)。また、cell cycle, DNA replication, mitotic cell cycleの辞書側の階層構造を上段に示している(73)。実験データではクラスターAに属するSTE7が他のAに属する遺伝子と異なり、cell cycle関連の用語と有意なネットワークを持たず、response to pheromone とネットワークを持つことから、実験ノイズが原因のミスクラスタリングであり、本来はクラスターBに属することが示唆される。アレイデータでは多数の遺伝子を対象とすることから、ミスクラスタリング候補となるネットワーク及び遺伝子は強調表示されることが望ましい。一方、YDR324はresponse to pheromoneとのネットワークは持たないが、その他の用語とも有意なネットワークを持たず、response to pheromoneに関わる新しい遺伝子である可能性が示唆される。

【0042】

クエリーは複数にすることによって、ネットワークの漏れを助けるという役目がある。特に十分な文献がない場合は、複数のクエリーにする効果大きい。例えば、CLF1ではDNA-replicationとの関係は文献から抽出できるが、cell cycleとの直接の関係は抽出できない。DNA replication とcell cycleの関係は文献から抽出できるので、この場合は問題

ないが、hemolysis と apoptosis など用語（概念）のように上下関係の概念であっても共起などで抽出し難い用語関係の場合は、関係を見落とすことになる。従って、apoptosis 関連をクエリー 1 に設定したい時は、クエリー 1 には apoptosis と hemolysis の両方を入れた方が、より漏れがなく確実に用語のネットワークを構築することができる。

【0043】

上述の実施例によれば、特に、連鎖解析、関連解析などで、疾患候補として出る領域には、遺伝子数が 100 を下回らないことが多く、これらの遺伝子情報を逐一論文を人が読み確かめていたのでは膨大な時間がかかる。また、専門外の分野の遺伝子を広く扱う場合、背景知識が欠如することとなり、2つの概念/用語の関連性を知らないがために、正解（疾患遺伝子）にたどりつけない可能性もある。この方法によって、短時間で候補遺伝子と疾患の関係を知ることができ、次に必要な実験に進むことができる。

【0044】

また、DNAアレイ、proteinアレイのデータにおいては、ノイズが多く含まれていることが知られており、発現データから精度を高く遺伝子のクラスタリングを行うことは容易ではない。この用語のネットワークを使うことにより、既に機能が知られている遺伝子については、ノイズが原因のミスクラスタリングの候補となる遺伝子を容易に見つけることができる。

【実施例 4】

【0045】

本実施例では、検索条件の指定をインタラクティブに行う実施例を、図 15 を用いて説明する。この例では、最初に第 1 と 2 のクエリーを結ぶネットワークの最長ステップ数を 4 と指定している。計算処理部ではデータ格納部から第 1 と 2 のクエリーに関係する情報をインタラクティブに抜き出しネットワークを構築し、そのノードの座標とエッジ情報を出力部に送り、ネットワーク-1 を表示する。ユーザはその結果を見て、予想した関係がでないと考えた場合、描画の最長ステップ数の変更を指定する。（図 15 中では 5 に変更）計算処理部で、再度ネットワークを計算し出力部でネットワーク-2 を表示する。再度の表示により余計なノードが画面に現れており、かつ、新たに 2 項関係を仮定するとどうなるかとユーザが考えた場合、そのノードを画面から削除し、新たに、2 用語間の関係を仮定して描画を指定する。計算処理部では、最初のクエリー情報に加え、新たに加わった 2 項関係に関わる情報をデータ処理部から取り出し、削除したノード情報を考慮しながらネットワークを計算し、出力部でネットワーク-3 を表示する。ノードの追加、削除、エッジ（2 項間関係）の追加、削除はネットワークの画面上、Java（登録商標）-アプレットの機能などを使って行ってもよいし、図 14 で示したテキスト形式で入力することも可能である。

【0046】

ここでは、計算処理部とデータ格納部がインタラクティブにデータのやりとりをしているが、データが十分に小さいときはこのシステムを立ち上げた時点でデータ格納部のデータを全て計算処理部のメモリー上に載せてしまい、同様の処理を行っても構わない。

【0047】

なお、第 1 のクエリーと第 2 のクエリーとを繋ぐ用語ネットワークの構築に際し、着目する組織で発現していない遺伝子/蛋白質を不使用とする設定を行えるようにしても良い。この設定を、インタラクティブに行えるようにしても、勿論良い。

【0048】

また、第 1 のクエリーと第 2 のクエリーとを繋ぐ用語ネットワークの構築に際し、図 14 の検索条件の指定 83 に示したように、用語の関連性の根拠となる記述についてその出典元となる学術誌のインパクトファクターの下限値を設定することができるようにし、その値以上の学術誌から抽出された用語間の関連性を使って、ネットワークを構成するようにしても良い。この設定を、インタラクティブに行えるようにしても、勿論良い。

【0049】

また、第 1 のクエリーと第 2 のクエリーとを繋ぐ用語ネットワークの構築に際し、用語

の関連性の根拠となる実験方法について、信頼性の低い大量データを生産する傾向のある実験手法（Yeast-two-hybridや質量分析法など）により発見された相互作用データを不使用とする設定することができるようにすると、ノイズを減らせるので良い。この設定を、インタラクティブに行えるようにしても、勿論良い。

【産業上の利用可能性】

【0050】

本発明は、実施例記載のバイオ情報検索の他、他のカテゴリ情報検索にも用いることができる。

【図面の簡単な説明】

【0051】

【図1】本発明における用語ネットワークシステムの実施の一例を示すシステム構成図。

【図2】図1におけるシステム利用の手順。

【図3】本発明におけるデータ格納装置において、格納すべきデータのうち自動抽出に関するデータの抽出手順。

【図4】本発明におけるデータ格納装置において格納する2項関係の情報例。図1では符号42に相当する。

【図5】本発明におけるデータ格納装置において格納する共起を計算するための用語の文献中の出現情報。図1では符号41に相当する。

【図6】本発明におけるデータ格納装置において格納する2項関係を構成する用語の付随情報。図1では符号44に相当する。

【図7】本発明におけるデータ格納装置において格納する2項関係の付随情報となる文献情報。図1では符号44に相当する。

【図8】本発明におけるデータ格納装置において格納する付随情報となる用語の階層関係。図1では符号44に相当する。

【図9】本発明におけるデータ表示例で、全ての用語を対象として記述した用語のネットワークを示す図。

【図10】本発明におけるデータ表示例で、上位の概念（用語）で記述した例。図8に相当する情報を使用する。

【図11】用語のネットワークにおけるスコアと連鎖解析によるロッドスコアを重ね合わせた例。図2でのステップ7に相当（符号25）

【図12】図11で、用語のネットワークにおけるスコアとロッドスコアが共に高かった部位での用語のネットワークの例。図2でのステップ8に相当（符号26）

【図13】DNAアレイの発現データによるクラスタリングとクラスタリングを構成する遺伝子とクエリー1間の用語のネットワークの描画例。図2でのステップ9に相当（符号29）

【図14】入力画面を表示した図。

【図15】実施例の検索のフローを示す図。

【符号の説明】

【0052】

1. 入出力装置/クエリー入力部
 11. クエリー1の入力部
 12. クエリー2の入力部
 13. 検索条件入力部
2. 中央計算処理装置
3. 入出力装置/表示部
4. データ格納装置
 41. 共起を計算するための文献中の用語の出現情報
 42. フレーズパターンで抽出した用語の2項関係
 43. 外部データベースから集めた2項関係と人手で抽出した2項関係などのその他の情

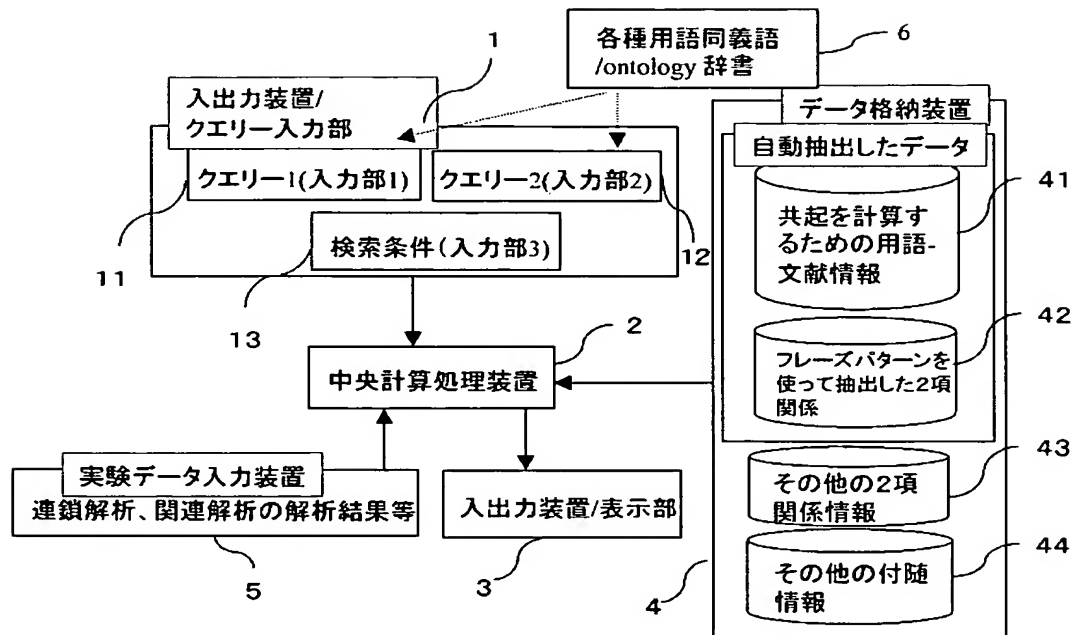
報

- 44. 2項関係を構成している用語の付随情報等
- 5. 実験データ入力装置
- 6. 各種同義語辞書
- 20. ステップ1: クエリー1の入力ステップ
- 21. ステップ3: クエリー2の入力ステップ
- 22. ステップ2: 連鎖解析、関連解析結果が存在するときにその結果を入力するステップ
- 23. ステップ2: アレイデータのクラスタリング結果が存在するときにその結果を入力するステップ
- 24. ステップ4: 検索条件の入力ステップ
- 25. ステップ7: 連鎖解析、関連解析の結果と用語のネットワークのスコアを重ねて表示するステップ
- 26. ステップ8: 疾患遺伝子の有力候補となる部分の用語のネットワークを表示するステップ
- 27. ステップ5: クエリー1の用語群とクエリー2の用語群を結ぶ用語のネットワークを表示するステップ
- 28. ステップ6: クエリー1とクエリー2を強く結ぶ用語を強調表示するステップ
- 29. ステップ9: アレイの発現データによる遺伝子クラスタリング結果と用語のネットワークを比較して、ミスクラスタリングと思われる遺伝子及びネットワークを提示するステップ
- 51. 2項関係の情報を抽出するための文献を収集するステップ
- 52. 遺伝子、蛋白質、化合物、機能用語などの用語を認識するステップ
- 全ての名詞句を用語として使う場合は、ここで構文解析や係り受け解析を行う
- 53. 構文解析、係り受け解析を行った後、文の構造解析を行うステップ
- 54. フレーズパターンにより2項関係を抽出するステップ
- 55. 文献中の用語とその出現位置をインデックス化するステップ
- 61. クエリー1とクエリー2に相当する遺伝子群を繋ぐ用語のネットワークを表示例する部分
- 62. 染色体上の遺伝子名を表示する部分
- 63. ロッドスコアと用語のスコアを合わせて表示する部分
- 71. クエリー1とクエリー2に相当するクラスタリングに使われた遺伝子群を用語のネットワークで繋ぐ部分
- 72. DNA アレイの発現データからの階層的クラスタリングを表示する部分
- 73. クエリー1の用語の階層構造を示す部分。

【書類名】 図面

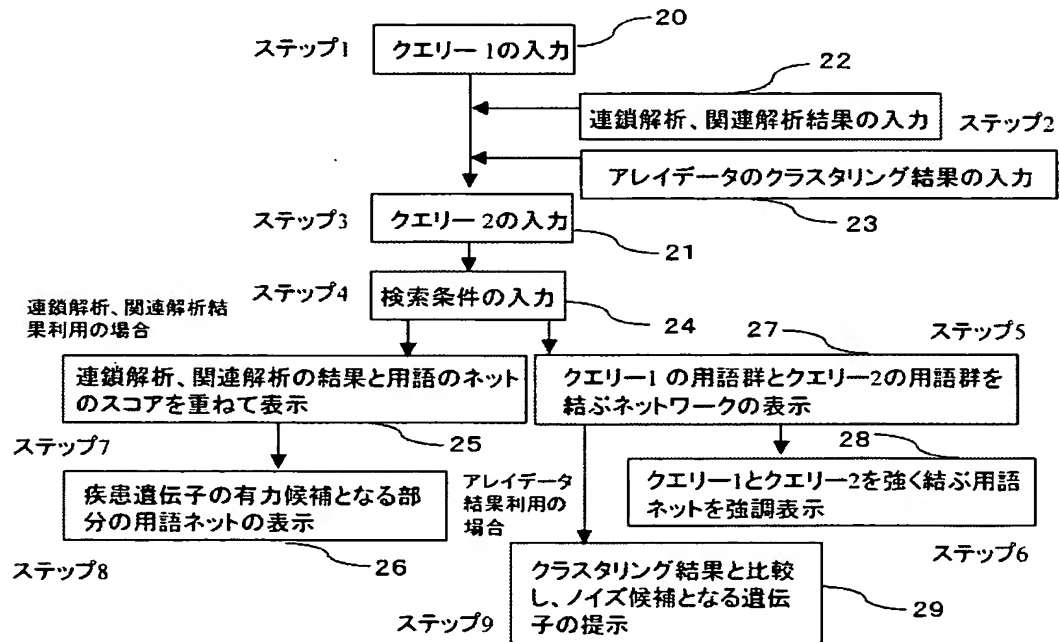
【図 1】

図 1

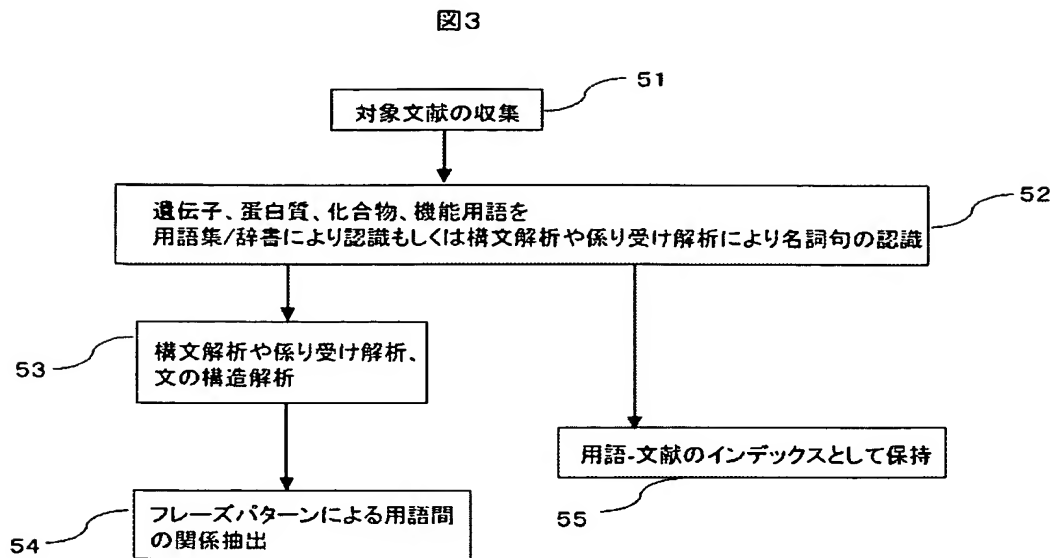


【図 2】

図 2



【図 3】



【図 4】

図4

protein-ID1 /concept1	protein-ID2 /concept2	信頼度	実験方法	生物種	文献-ID
GSC004154	GSC004160	0.95	Yeast-two hybrid, 質量分析	<i>S. cerevisiae</i>	1, 2
GSC004168	GSC004154	0.9	Yeast-two hybrid, 質量分析	<i>S. cerevisiae</i>	3
cell_death	apoptosis	0.8	—	<i>H. sapiens</i>	5
GHS001223	apoptosis	0.9	—	<i>H. sapiens</i>	4
GHS001223	BH2 domain	0.5	—	<i>H. sapiens</i>	6
GHS001223	GHS016577	0.9	—	<i>H. sapiens</i>	7
...	...				

【図 5】

図5

protein-ID1/concept1	PUBMED-ID	出現頻度	文献中の出現位置 (バイト)	文献-ID
GSC004154	12909353	3	305, 777, 930	101
GSC004154	12867033	1	922	102
GSC004154	12827445	1	417	103
GSC004154	12808050	2	607, 1272	104
...	...			

【図 6】

図6

ID	名前	生物種	細胞内局在性	配列情報	ドメイン情報	発現情報
GSC004154	ste11	S. cerevisiae	cytoplasmic	GCE000836:1e-22, GCE011584:5e-21	Ser/Thr protein kinase domain	
GSC004160	ste20	S. cerevisiae	cytoplasmic	GCE000836:4e-32, GCE000678:1e-38	Ser/Thr protein kinase domain	
GSC004168	ste7	S. cerevisiae	cytoplasmic	GCE000822:9e-45, GCE000667:1e-31	Ser/Thr protein kinase domain	
GHS012062	MAPK1	H. sapiens	cytoplasmic	GCE000884:1e-164,	Ser/Thr protein kinase domain	neuroblastoma cot, lymph, brain.
GCE011584	K06H7.1	C. elegans	-	GSC004154: 5e-21		
GHS001223	BCL-2	H. sapiens	mitochondrial membrane, intracellular membrane of the nuclear envelope, the endoplasmic reticulum.		BH4 domain BCL-2 domain	blood, kidney, skeletal, lymphocyte,...
GHS016577	PSEN1	H. sapiens	Integral membrane protein, Golgi and endoplasmic reticulum	GCE001332:7e-95 GCE000580:8e-37	presenilin 1 domain	brain, skin
...		...				

【図 7】

図 7

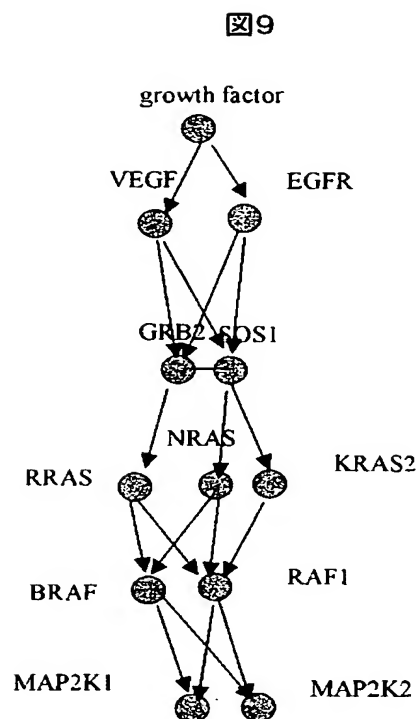
文献ID	impact factor	pubmedID	journal	抽出センテンス
1	8.4	108837245	Curr Biol	Our results suggest that . in response to multiple extracellular signals , phosphorylation of Ste11p by Ste20p removes an amino-terminal inhibitory domain , leading to activation of the Ste11 protein kinase .
2	8.4	10837245	Curr. Biol.	Ste20p phosphorylated Ste11p on Ser302 and/or Ser306 and Thr307 in yeast , residues that are conserved in MEKKs of other organisms .
3	10.8	8052657	PNAS	Interaction between STE7 and STE11 is bridged by STE5 . suggesting the formation of a multiprotein complex .
4	4.9	12769779	Curr Med Chem	Bcl2 proteins are key mediators of the process of apoptosis and ligands to these family of proteins have been described using modern combinatorial.
...	...			

【図 8】

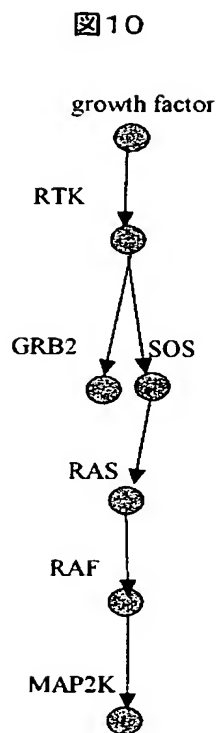
図 8

用語(概念)	上位の用語(概念)
ste7	MAPK
MAP2K1	MAPK2K
MAP2KK2	MAPK2K
ste11	MAP2K
ste20	MAP3K
MAPK	ser/thr kinase
EGFR	RTK
VEGF	RTK
RRAS	RAS
NRAS	RAS
KRAS2	RAS
BRAF	RAF
RAF1	RAF
...	

【図 9】

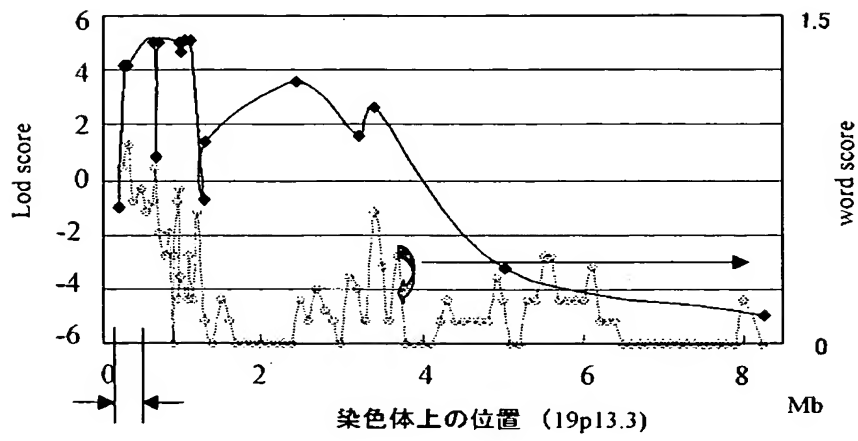


【図 10】

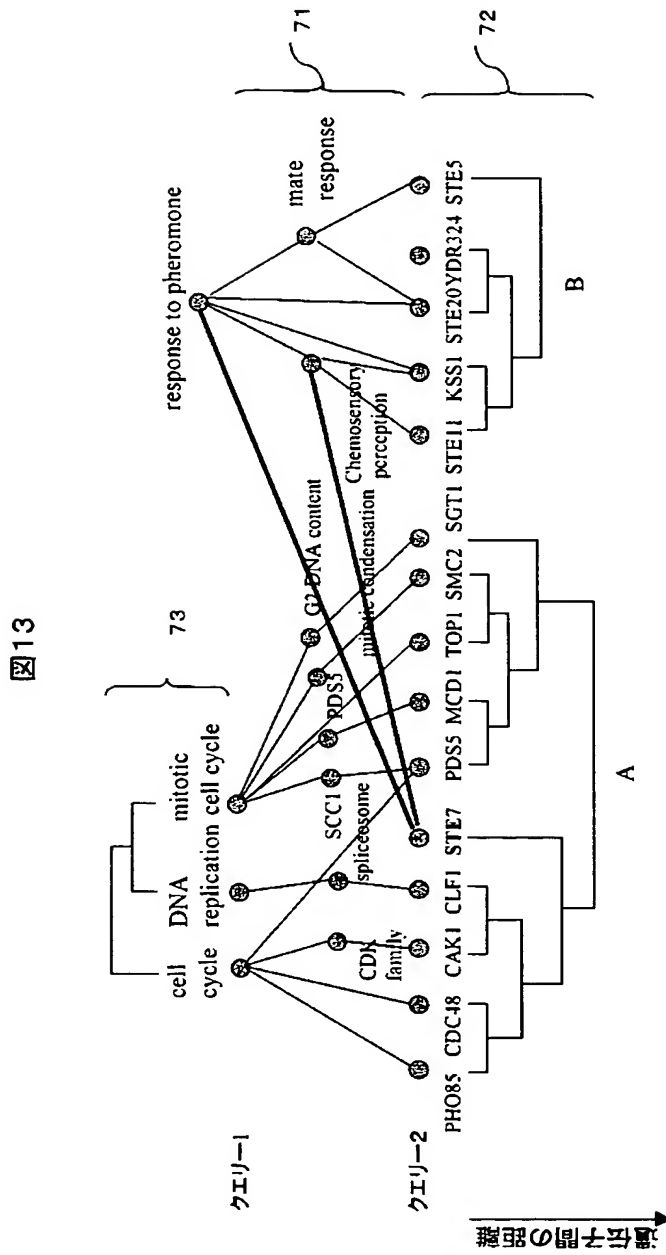


【図 11】

図 11

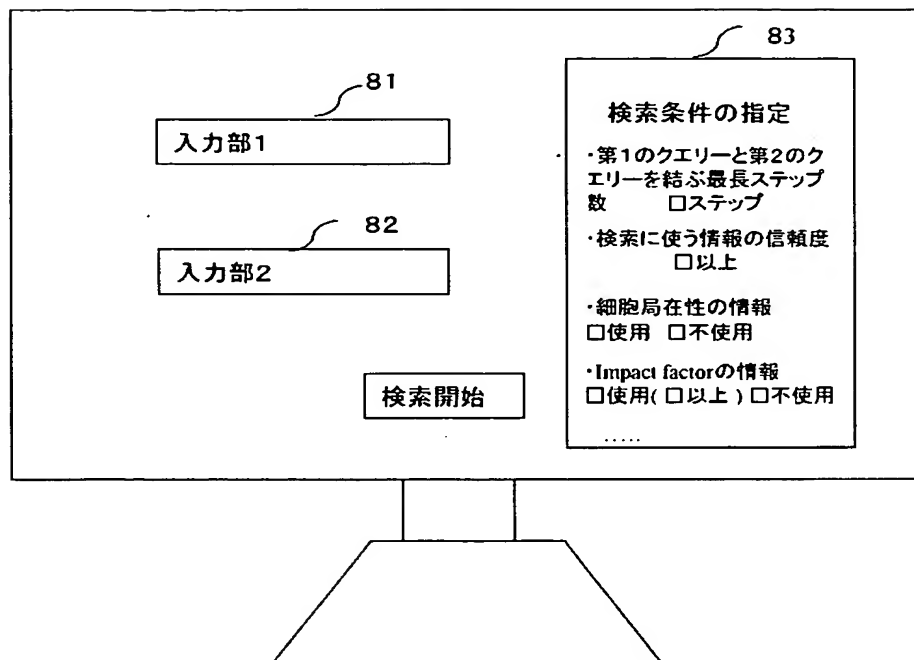


【図13】

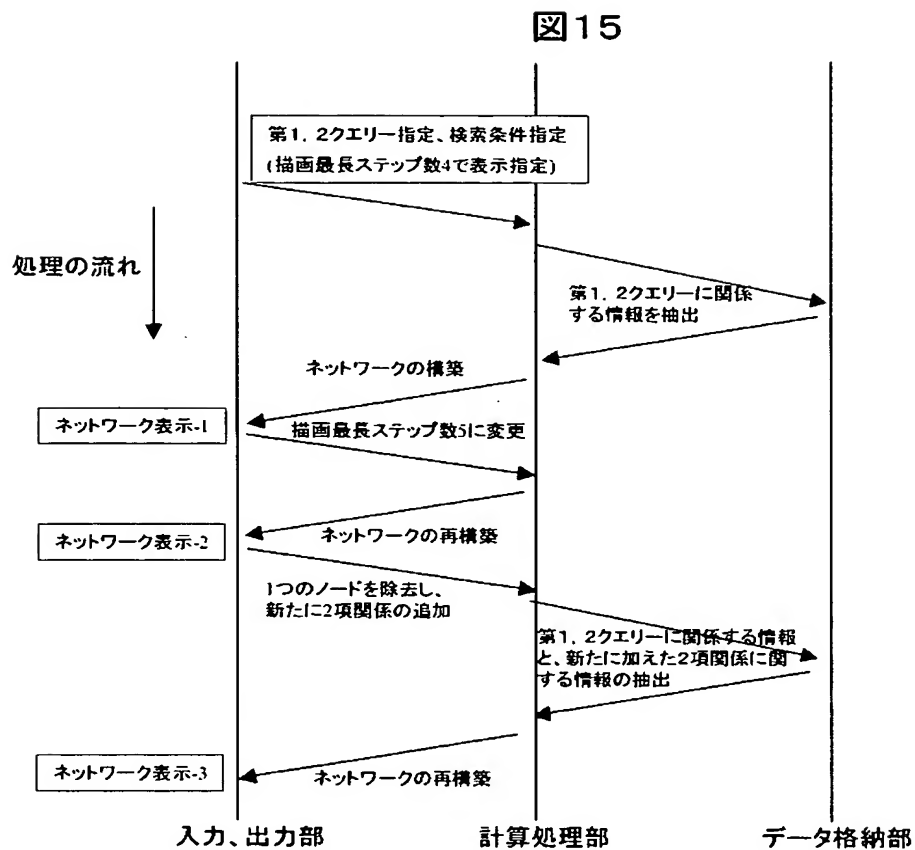


【図 14】

図14



【図 15】



【書類名】 要約書

【要約】

【課題】 遺伝子に関する実験的に新たに得られたデータと文献から得られたデータ及びインターネット等を介して得られるデータを統合することにより、新しい知見を得ること。

【解決手段】 データ格納装置 4 に蓄積した遺伝子、化合物、疾患、遺伝子の機能等の用語間の関連性情報を用いて、クエリー入力部 1 で指定されたクエリー 1-1 1 と クエリー 2-1 2 の間を結ぶ用語のネットワークを再構成し、表示部 4 により表示することにより、クエリー 1, クエリー 2 を関連付ける用語を表示する。

【効果】 クエリー 1 とクエリー 2 がいかんして関連付けられるかという知見をユーザに与える。

【選択図】 図 1

認定・付加情報

特許出願の番号	特願 2 0 0 3 - 3 5 3 0 9 7
受付番号	5 0 3 0 1 6 9 8 9 0 2
書類名	特許願
担当官	第七担当上席 0 0 9 6
作成日	平成 1 5 年 1 0 月 1 5 日

< 認定情報・付加情報 >

【提出日】	平成 1 5 年 1 0 月 1 4 日
-------	----------------------

特願 2 0 0 3 - 3 5 3 0 9 7

出 願 人 履 歴 情 報

識別番号

[0 0 0 0 0 5 1 0 8]

1. 変更年月日

1 9 9 0 年 8 月 3 1 日

[変更理由]

新規登録

住 所

東京都千代田区神田駿河台 4 丁目 6 番地

氏 名

株式会社日立製作所